Project Proposal **Evan Nichols, Keith Monaghan, Casey Sader**
EECS 731 9-15-16
Prof. Beckage

1 Abstract

Vocabulary development is the process by which people acquire words. The Stanford Wordbank (wordbank.stanford.edu) provides extensive data over vocabulary growth curves across a variety of demographic groups, allowing researchers to compare the effects of different variables (e.g. age in months, birth order, ethnicity, gender, mother?s education) on development. We are interested in seeing which models can best capture the acquisition of language in children and what roles these variables play in the speed of language acquisition. We will reserve a portion of the dataset for testing and validating each model. After creating and testing the models, we can explore different data visualization tools to better illustrate our findings.

2 Abstract

We plan to compare and contrast the effectiveness of different algorithms on modeling children's language development. We will consider algorithms from the following list, all available through the Python scikit-learn library: Naive Bayes, Logistic Regression, Random Forest. We will also explore the ability of Deep Neural Networks (using Caffe or TensorFlow) to model our data.

We will use the variables available through the Stanford Wordbank dataset: age in months, birth order, ethnicity, gender and mother?s education level. We will use the above models to identify the variables most important to language acquisition, and visualize our findings using Python?s matplotlib library. We anticipate the mother's education level will have the largest impact (excluding age) on production and comprehension.

To test the models, we will randomly sample the dataset and withhold a portion of the data from the training set to validate and evaluate the performance of the model. Scikit-learn provides the tools necessary to evaluate the predictions generated by the model using metrics like accuracy, precision, f1 score, recall, and others which can tell us where the model is performing well and where it is making the most mistakes. We anticipate that deep neural networks will perform the best.

After we complete experimentation, we will use what we learned to visualize portions of the dataset to highlight key features of the dataset. We hope to create graphs that intuitively demonstrate the importance of certain variables on language acquisition. We also hope to visualize the results of our experimentation using different machine learning tools to model the data.

3 Timeline

- September
  - Project Proposal
  - Research and document related work

  - Develop and refine hypothesis

- October
  - Preprocess dataset
  - Finalize methods
  - Finalize hypothesis

- November
  - Finish experiments
  - Create Data Visualizations
  - Begin work on report and final presentations

- December
  - Present Final Presentation
  - Submit Final Report

4 Related Work

A 2015 study discovered strong correlations between the order of words learned across 25 different languages. Examining data from 299 words (158 nouns and 141 verbs), the researchers found the Age of Acquisition (AoA) ratings in all language were significantly correlated (Spearman's rho, adjusted for split-half reliabilities) [2]. The highest correlation was found between Polish and Slovak (r = .96) and British and South African English (r = .91), and the lowest between Hungarian and Italian (r = .62) and Irish and Hebrew (r = .65). This study seems to indicate a strong relationship between lexical similarity and AoA.

A 2011 study by Norbert Schady illustrated the schooling and vocabulary levels of mothers were strong predictors of the cognitive development of young children [1]. Analyzing longitudinal data from 6 rural provinces in Ecuador, including the parents' education and vocabulary scores, household wealth and total per capita expenditures. Schady illustrated strong correlations between a child's cognitive development and the mother's years of schooling and vocabulary. Increasing the maternal vocabulary by 1 SD increased children's scores on the vocabulary, memory and visual integration tests by 0.24, 0.24 and 0.20, respectively. The study seems illustrates a rather intuitive observation: children tend to be more educated given they have wealthy and educated parents.

References

[1] S. Luniewska M. Haman E. Armon-Lotem. Ratings of age of acquisition of 299 words across 25 languages: Is there a cross-linguistic order of words? *Behavior Research Methods*, 48(3):1154–1177, sep 2016.

[2] N. Schady. Parents education, mothers vocabulary, and cognitive development in early childhood: Longitudinal evidence from ecuador. *American Journal of Public Health*, 101(12):2299–2307, dec 2011. 10.2105/AJPH.2011.300253.